Published online 2 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D475–D479***
doi:10.1093/nar/gkm884

The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata

Konstantinos Liolios[1], Konstantinos Mavromatis[2], Nektarios Tavernarakis[3] and Nikos C. Kyrpides[2,*]

[1]University of Chicago, Department of Medicine, Chicago, [2]Genome Biology Program, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA and [3]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology, Heraklion, Crete, Greece

Received September 18, 2007; Revised and Accepted October 1, 2007

ABSTRACT

The Genomes On Line Database (GOLD) is a comprehensive resource that provides information on genome and metagenome projects worldwide. Complete and ongoing projects and their associated metadata can be accessed in GOLD through pre-computed lists and a search page. As of September 2007, GOLD contains information on more than 2900 sequencing projects, out of which 639 have been completed and their sequence data deposited in the public databases. GOLD continues to expand with the goal of providing metadata information related to the projects and the organisms/environments towards the Minimum Information about a Genome Sequence' (MIGS) guideline. GOLD is available at http://www.genomesonline.org and has a mirror site at the Institute of Molecular Biology and Biotechnology, Crete, Greece at http://gold.imbb.forth.gr/

HISTORY AND GROWTH

Since its inception in 1997, GOLD (1–3) has continuously monitored genome sequencing projects worldwide and has provided the community with a unique centralized resource that integrates diverse information related to Archaea, Bacteria, Eukaryotic and more recently Metagenomic sequencing projects.

In contrast to what was anticipated in the previous report of the database 2 years ago (3), the total number of projects has not yet doubled, with a total of 2905 projects recorded as of September 2007 (compared to 1575 in September 2005). However, if only the archaeal and bacterial projects are considered, then the total number of recorded projects is 1950, only 36 short of double the number of such recorded projects 2 years ago. The advent of new sequencing technology platforms, such as pyrosequencing (4), has significantly contributed to the increase in the number of new microbial sequencing projects. In fact, 134 GOLD projects are now reported using 454 technology platform as part of the Whole Genome Sequencing (WGS) project.

Two major large-scale microbial genome sequencing programs that have been launched during the last 2 years account for the majority of the reported 454 sequencing projects. The first program is the Human Gut Microbiome Initiative (HGMI) (http://genome.wustl.edu/hgm/HGM_frontpage.cgi) at the Genome Sequencing Center of the Washington University in St. Louis. This program aims to provide deep draft genome sequences for 100 cultured representatives of the phylogenetic diversity documented by 16S rRNA surveys of the human gut microbiota. From these, the sequencing for 45 organisms is already in progress and the information available in GOLD (the list is available through the search page with the term 'Human gut microbiome' as the Relevance search field). The second program is the Genomic Encyclopedia of Bacteria and Archaea (GEBA), launched in May 2007 by the Department of Energy (DOE)—Joint Genome Institute (JGI) (http://www.jgi.doe.gov/programs/GEBA/index.html). GEBA aims the systematic filling in the sequencing gaps along the bacterial and archaeal branches of the tree of life and represents the first systematic attempt to use the tree of life itself as a guide for sequencing target selection. To test the feasibility of such a large-scale project, DOE-JGI has initiated a pilot project to sequence 100 bacterial and archaeal organisms based on their phylogenetic position in the tree of life. The GEBA pilot project is carried out in collaboration with the German Resource Centre for Biological Material (DSMZ) (http://www.dsmz.de/), which provides the DNA for the selected organisms. As of September 2007, 79 GEBA projects are reported on GOLD (the list is available

*To whom correspondence should be addressed. Tel: 925 296 5718; Fax: 925 296 5850; Email: nckyrpides@lbl.gov

through the search page with the term 'GEBA' as the Relevance search field).

In addition to the HGMI and GEBA programs, several National and International efforts for systematic exploration of the Biodiversity have been initiated during the last few years, and are also expected to contribute to a significant increase in sequencing projects. These efforts include the MikroBioKosmos initiative in Greece (http://www.mikrobiokosmos.org/), the Australian Genome Alliance (http://www.genomealliance.org.au/), the Biodiversity Research Initiative in Germany (http://www.dfg.de/en/news/press_releases/2006/press_release_2006_25.html), the National BioResource project in Japan (http://www.nbrp.jp/index.jsp) and the International Census for Marine Microbes (ICoMM) (http://www.coml.org/descrip/icomm.htm), among others.

During the last 2 years, GOLD has been extended in terms of its project tracking capability to record the emerging metagenome projects and to comprehensively capture metadata for all projects. These extensions are further discussed below.

In order to facilitate more efficient project tracking, the sequencing centers and the community at large are highly encouraged to register the sequencing projects in GOLD. Further, in order to facilitate cross-reference between different resources, it is recommended that the genome submission files, should also include the corresponding GOLD ID.

CURRENT STATUS OF GOLD

Published complete genomes

As of September 2007, GOLD records 639 completed genome projects, which is more than double the number reported 2 years ago (3). These projects have their complete sequence deposited into the public archival sequence databases such as GenBank (5), EMBL (6) and DDBJ (7). Some of these projects do not have an associated publication since often submitters release their sequence data to the community prior of preparing or submitting a publication. This approach has increased significantly the speed of releasing complete genome sequences to the benefit of the scientific community. From the total of 639 complete and published genome projects, 527 are bacterial, 47 are archaeal and 65 are eukaryotic. For 56 of the 65 eukaryotic projects reported as complete, the sequence status is reported as Quality Draft (information available in the data download file).

Ongoing genome projects

In addition to the complete projects, there are currently 2158 ongoing sequencing projects. 1328 of those are bacterial, 59 archaeal and 771 eukaryotic projects. The latter include 271 EST projects, 74 projects that focus on specific genomic regions or constitute general genome surveys, and 426 whole-genome sequencing projects. These can be retrieved using GOLD's search engine, selecting 'EST' or 'Genome-Regions' or 'Genome-Survey' for the Type field.

From the 2158 ongoing projects, 125 are also considered complete at this point, that is the sequencing phase has been completed but the data are not yet submitted to the public genome sequence archives and 513 have already a draft version available. These can be retrieved using GOLD's search engine through the Status field.

A number of the reported projects (either complete or ongoing) are proprietary and their data may never be publicly released. There are currently 86 such projects reported on GOLD and they can be retrieved by selecting 'Proprietary' for the Availability field of the Search page. Usually, only the information for the sequencing project itself has been made available in these cases.

Metagenome projects

During the last 2 years a constantly growing number of metagenomic projects have being initiated, and the expectation is that their number will keep on growing as the sequencing technology improves. GOLD currently reports 108 distinct metagenome projects, 25 of which are considered complete. For GOLD, a metagenome project is considered complete when the data are deposited in the public sequence archives and a paper describing the project is published. The organization, structure and presentation of the metagenome data is described in more detail below.

Metadata

Two types of metadata are provided by GOLD: (i) project metadata and (ii) organism/environment metadata. The current status of the different fields and the number of projects with associated data for each of the corresponding fields, are shown in Table 1. Some of the metadata fields are populated with information for all or most of the projects, while other fields (particularly newer ones such as the pH), are yet to be curated for the majority of the projects. While the number of different metadata types will be gradually expanding, the current list is already used in the context of microbial comparative analysis systems such as the Integrated Microbial Genomes IMG (8) and IMG/M (9) systems, the xBASE database (10) and the wireless genome information WiGID database (11).

NEW DEVELOPMENTS

Organization of metagenomic projects

The semantics, organization and presentation of metagenome projects, are still evolving. Given the fundamental differences that they have compared to the isolate genome projects, in most cases new storing, organization and presentation methods need to be developed. Some of the main challenges in tracking and reporting metagenome projects include: (i) definition of a metagenome project, (ii) standardized description of the project name, (iii) classification of metagenome projects, (iv) capturing and displaying the distinct samples associated with a project, (v) capturing and displaying the phylogenetic distribution of the organisms in every sample, (vi) capturing and displaying the metadata for individual samples as well as

Table 1. Metadata types available from GOLD

Project metadata fields	Number of projects	Organism/ environment metadata	Number of projects
1. GOLD Project ID	2905	1. Domain	2905
2. GCAT ID	2905	2. Phylum	2905
3. NCBI Project ID	1903	3. Class	2905
4. IMG OID	829	4. Order	2905
5. Sequencing method	797	5. Family	2905
6. Sequencing coverage	401	6. Genus	2905
7. Project type	2905	7. Species	2905
8. Sequencing status	2905	8. Strain	2113
9. Project status	1375	9. Serovar	177
10. Country	2905	10. Taxon ID	2806
11. Availability	2905	11. StrainInfo ID	320
12. Sequencing center	2896	12. Greengenes ID	707
13. Project relevance	2241	13. Culture Collection ID	595
14. Funding center	2108	14. Size	1717
15. Sequence data	1160	15. Gene number	991
16. Database	1983	16. Chromosome number	793
17. Publication	448	17. Plasmid number	777
18. Release date	664	18. GC%	1184
19. Contact name	2158	19. Phenotype	2123
20. Contact email	2150	20. Habitat	1962
		21. Disease	983
		22. Temperature	626
		23. pH	69
		24. Isolation	1023
		25. Symbiont	122

for the entire project and (vii) lack of availability of standardized metadata ontology.

GOLD will gradually address all problems over the next several releases by expanding towards the MIGS/MIMS guideline (12). The current version of GOLD has addressed the first three problems mentioned above:

(i) Definition of a metagenome project: there is already a lot of confusion and ambiguity regarding the definition of a metagenome project. Sometimes individual samples that are part of a study constitute a separate project, while in other cases, all the samples are grouped into a single project. In order to avoid such discrepancies, a mategenome project in GOLD is defined as a single study with related samples presented as part of the same project. For example the project Gm00071 (http://genomesonline.org/GOLD_CARDS/Gm00071.html), has five samples.

(ii) Standardized description of the project name: this is a major problem for the emerging metagenome projects, with the same study (project) often associated with different names in various resources. As the number of projects grows it will be increasingly difficult to track projects across different databases, without a standardized naming convention. In order to address this problem, GOLD has implemented in the metagenome project names the genus–species–strain structure employed for isolate genomes. Accordingly, each metagenome project name consists of up to four parts: (a) Project Habitat (equivalent to Genus level), which describes the habitat of the community, e.g. Air, Gut, Endophytic, Soil, Wastewater, Fossil, Marine, etc. (b) Project Community (equivalent to species), which describes the nature of the community under study, e.g. microbial, fungal, viral, etc. (c) Project Location (equivalent to strain), which describes the location of the community, e.g. Human, New York, Bioreactor, etc. and (d) Project Identifier, which describes the specific type (identifier) of the community, e.g. lean and obese, adults, thermal gradient, etc. This naming convention will help avoiding cases where one project would be named New York Air, while another will be named Air from New York or air from Singapore. This project naming structure is essentially employing a combination of specific metadata fields in order to synthesize the complete project name, and will also help grouping, sorting and searching projects based on habitat, community, location and various identifiers.

(iii) Classification of the projects: similar to the two problems described above, a classification schema analogous to the Taxonomic classification available for isolate organisms, does not yet exist for metagenomes. In order to address this problem, the current version of GOLD organizes metagenome projects in three main categories: (a) Environmental (e.g. Environmental-Air, Environmental-Marine, etc.), (b) Endobiotic (e.g. Endobiotic-Human, Endobiotic-Plants, etc.) and (c) Synthetic (e.g. Synthetic-Simulated, Synthetic-Bioreactor, etc.). The GOLD classification for Metagenome projects is presented in the Information field in the metagenome project list and also available through the Search page, under Metagenome Classification.

New data fields

In addition of tracking metagenome projects, since the last report (3), a number of additional data fields have been added to the database, both in the project tables, and its search engine. These include the fields (i) Country, which displays the name of the countries that have genome project (all the projects are currently distributed across 31 countries including a few multinational efforts), (ii) Sequencing method, which denotes if 454 or other technologies are used for sequencing; (iii) Sequencing depth; (iv) pH; (v) Temperature, (vi) Project Status, which distinguishes the completion of sequencing versus the completion of the project and (vii) Metagenome Samples as a separate field for each of the metagenome projects. In the future, these will be further developed to allow the capturing of individual metadata for each of the samples in addition to the metadata for the entire project.

New pages

A number of new pages have been added. These include: (i) GOLD CARD pages for every project, which is available from the link of every GOLD_STAMP ID. The information in every one of these pages is organized into three tables: (a) Organism information, (b) Genome

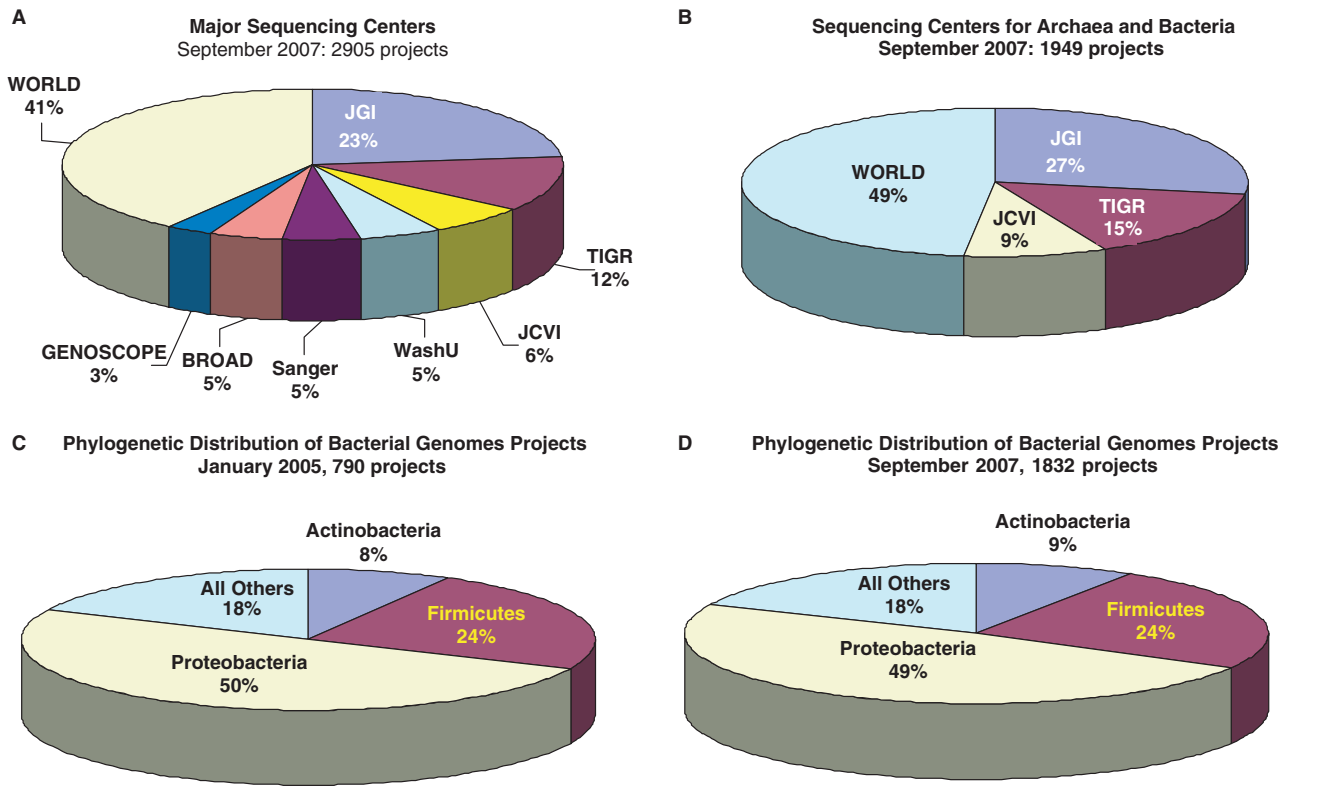

Figure 1. Statistical information available in GOLD. (**A**) Distribution of the 2995 genome projects across the major sequencing centers. Abbreviations are for, JGI: Joint Genome Institute, TIGR: The Institute for Genome Research, JCVI: J. Craig Venter Institute, WashU: Washington University and WORLD: all other sequencing centers. (**B**) Distribution of the 1949 bacterial and archaeal genome projects across the major sequencing centers. (**C**) Phylogenetic distribution of the 790 bacterial genome projects in January 2005. (**D**) Phylogenetic distribution of the 1832 bacterial genome projects in September 2007.

project information and (c) External links. Future developments here will include expanding the information and reorganizing the structure of the three tables closer to the structure shown in Table 1; (ii) Taxonomic Tree of the projects. Here, the NCBI taxonomy is used to display the number of GOLD sequencing projects down to the Genus level. This is quite helpful in identifying taxonomic groups that are not yet covered from sequencing projects.

Data availability and cross-referencing

All GOLD data are available according to the Creative Commons License of Attribution-NonCommercial-ShareAlike (http://creativecommons.org/licenses/by-nc-sa/2.5/). Most of the data can now be downloaded to an excel file, in order to facilitate data distribution and wider use. A number of additional data fields that are not available either in the project tables or in the search page are now available directly for download. These include (i) GreenGenes (13); (ii) StrainInfo (14); (iii) GCAT (http://gensc.sf.net) and (iv) IMG (8) and IMG/M (9) IDs. These identifiers provide cross-referencing between the resources mentioned above with those from NCBI such as RefSeq (15), Entrez Project and Taxonomy IDs. Additional fields in this file include the NCBI Taxonomic levels of Superkingdom, Phylum, Class, Order, Family, Genus and Species.

Other data available for download include a regularly updated statistical data file, which is accessible from the Statistics link of the front page (see below).

OVERVIEW STATISTICS

Although several different types of statistics, related to each of the data fields, can be derived from the users at any point using the search engine, or the available for download data, GOLD also provides graphical overviews for specific data types. These are provided through the link 'Gold Statistics' available on the home page of the database, and include the following data types

Sequencing centers

More than half of the 2900 currently available sequencing projects on GOLD are distributed among only four major sequencing centers (since TIGR and the Venter Institute have recently merged). When only the archaeal and bacterial projects are taken into account, two sequencing centers seem to carry more than half of the world's production. These are the Joint Genome Institute (JGI) and the Venter Institute (JCVI) with TIGR. On top of the list in both cases is the JGI, which is the Department of Energy (DOE) sequencing facility with 23% and 27% of world's production respectively (Figure 1). This is based on the number of unique individual projects, and do not

correspond directly to the actual size of the project or the number of sequenced bases that is harder to monitor.

Phylogenetic distribution

The sampling bias towards only three major bacterial lineages (Proteobacteria, Firmicutes and Actinobacteria) continues to persist despite the large increase in sequencing projects as was previously reported (3). As shown on Figure 1, even though the number of Bacterial genome sequencing projects has increased 2.3-fold over the last 2.5 years, the percentage of the three major lineages remains almost entirely unchanged. The development of novel methods that bypass the major restriction of culturing the organism for sequencing (16,17) will hopefully alleviate this bias.

DATABASE AVAILABILITY

GOLD can be accessed at http://www.genomesonline.org/ Further comments and feedback are welcome at mail@genomesonline.org.

ACKNOWLEDGEMENTS

GOLD has been maintained and developed mostly based on the volunteer work of its small team. We are grateful to all the colleagues who kindly provide information for the more accurate monitoring of the genome projects. The support of Tatiana Drakakis and Rashida Lathan, and the continuous contributions of Philip Hugenholtz, Tomer Altman, Krishna Palaniappan and Victor Markowitz, are especially acknowledged. The list of all contributors is available at: http://www.genomesonline.org/acknowledgments.html. The work presented in this article was partially supported by the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, US Department of Energy under Contract No. DE-AC03-76SF00098. Funding to pay the Open Access publication charges for this article was provided by the Department of Energy Joint Genome Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Kyrpides,N. (1999) Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, **15**, 773–774.
2. Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes Online Database (GOLD): a monitor pf genome projects world-wide. *Nucleic Acids Res.*, **29**, 126.
3. Liolios,K., Tavernarakis,N., Hugenholtz,P. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects world-wide. *Nucleic Acids Res.*, **34**, D332–D334.
4. Diggle,M.A. and Clarke,S.C. (2004) Pyrosequencing: sequence typing at the speed of light. *Mol. Biotechnol.*, **28**, 129–137.
5. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) GenBank. *Nucleic Acids Res.*, **35**, D21–D25.
6. Kulikova,T., Akhtar,R., Aldebert,P., Althorpe,N., Andersson,M., Baldwin,A., Bates,K., Bhattacharyya,S., Bower,L. *et al.* (2007) EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res.*, **35**, D16–D20.
7. Okubo,K., Sugawara,H., Gojobori,T. and Tateno,Y. (2006) DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res.*, **34**, D6–D9.
8. Markowitz,V.M., Korzeniewski,F., Palaniappan,K., Szeto,E., Werner,G., Padki,A., Zhao,X., Dubchak,I., Hugenholtz,P. *et al.* (2006) The Integrated Microbial Genomes (IMG) system. *Nucleic Acids Res.*, **34**, D344–D348.
9. Markowitz,V.M., Ivanova,N., Palaniappan,K., Szeto,E., Korzeniewski,F., Lykidis,A., Anderson,I., Mavromatis,K., Kunin,V. *et al.* (2006) An Experimental Metagenome Data Management and Analysis System. *Bioinformatics*, **22**, 359–367.
10. Chaudhuri,R.R. and Pallen,M.J. (2006) xBASE, a collection of online databases for bacterial comparative genomics. *Nucleic Acids Res.*, **34**, D335–D337.
11. Ursing,B.M. (2003) WiGID: wireless genome information database. *Bioinformatics*, **19**, 439–440.
12. Field,D., Garrity,G., Gray,T., Morrison,N., Selengut,J., Sterk,P., Tatusova,T., Thomson,N., Allen,M.J. *et al.* (2007) Towards a richer description of our complete collection of genomes and metagenomes: the 'Minimum Information about a Genome Sequence' (MIGS) specification. *Nat. Biotechnol.*, In press.
13. DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
14. Dawyndt,P., Dedeurwaerdere,T. and Swings,J. (2007) Exploring and exploiting microbiological commons: contributions of bioinformatics and intellectual property rights in sharing biological information. *Int. Soc. Sci. J.*, **188**, 258.
15. Pruitt,K.D., Tatusiova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence databse of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
16. Marcy,Y., Ouverney,C., Bik,E.M., Lösekann,T., Ivanova,N., Martin,H.G., Szeto,E., Platt,D., Hugenholtz,P. *et al.* (2007) Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl Acad. Sci. USA*, **104**, 11889–11894.
17. Podar,M., Abulencia,C.B., Walcher,M., Hutchison,D., Zengler,K., Garcia,J.A., Holland,T., Cotton,D., Hauser,L. *et al.* (2007) Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.*, **73**, 3205–3214.